Handling Systematic Uncertainties and Combined Source Analyses for Atmospheric Cherenkov Telescopes

Hugh Dickinson[a,*], and Jan Conrad[a]

[a]Oskar Klein Centre, Department of Physics, Stockholm University, Albanova University Center, SE-10691 Stockholm, Sweden

Abstract

In response to an increasing availability of statistically rich observational data sets, the performance and applicability of traditional Atmospheric Cherenkov Telescope analyses in the regime of systematically dominated measurement uncertainties is examined. In particular, the effect of systematic uncertainties affecting the relative normalisation of fiducial ON and OFF-source sampling regions - often denoted as α - is investigated using combined source analysis as a representative example case. The traditional summation of accumulated ON and OFF-source event counts is found to perform sub-optimally in the studied contexts and requires careful calibration to correct for unexpected and potentially misleading statistical behaviour. More specifically, failure to recognise and correct for erroneous estimates of α is found to produce substantial overestimates of the combined population significance which worsen with increasing target multiplicity. An alternative *joint likelihood* technique is introduced, which is designed to treat systematic uncertainties in a uniform and statistically robust manner. This alternate method is shown to yield dramatically enhanced performance and reliability with respect to the more traditional approach.

Keywords: Gamma-ray astronomy, Systematic Uncertainties, Combined source analysis, Stacking analysis, Data Stacking, Statistical analysis, Likelihood analysis

1. Introduction

Recent developments involving Imaging Atmospheric Cherenkov Telescopes (IACTs) have revolutionised the field of Very High Energy (VHE) γ-ray astronomy. Indeed, observational data obtained using the current generation of stereoscopic arrays have contributed over 100 new sources to the catalogue of confirmed GeV-TeV emitters[1].

---

*Corresponding author
  Email address: hugh.dickinson@fysik.su.se (Hugh Dickinson)
  [1]TeVCat: http://tevcat.uchicago.edu

Notwithstanding these impressive achievements, the motivation for astronomers to continue identifying and categorising additional γ-ray sources and source populations remains undiminished. The initial source detections are likely to be dominated by luminous or nearby γ-ray emitters, implying a requirement for subsequent observations to target progressively fainter source populations. Accordingly, the development of analytical techniques which enable maximal exploitation of the available instrumental sensitivity will become increasingly important as the life-cycles of the various IACT arrays unfold. This emerging philosophy is exemplified by the recent work of Klepser (2011) which employs accurate knowledge of the instrumental response to enhance the performance and reliability of an established technique for source detection originally developed by Li and Ma (1983).

Until recently, the inherent faintness of astrophysical VHE γ-ray fluxes has resulted in statistically limited signal measurements with uncertainties that are dominated by Poissonian shot noise. Accordingly, many traditional IACT analyses have justifiably neglected the existence of previously sub-dominant systematic effects. In contrast, successful source detection at the limits of instrumental sensitivity requires deep observations and the accumulation of rich data sets with abundant event statistics. In this new regime, the relative contribution of statistical fluctuations to the overall error budget is suppressed with measurement accuracy ultimately becoming systematically limited.

This work investigates methods for the appropriate statistical treatment of systematic uncertainties affecting IACT observational data and their subsequent analysis. In Cherenkov astronomy, the synthesis of large data sets typically involves aggregation of temporally sparse event statistics which correspond to distinct values of various observational parameters, each of which can introduce distinct, observation-specific systematic biases. Accordingly, analyses that are designed to ameliorate the impact of irreducible systematic effects must incorporate sufficient flexibility to successfully model this inter-observational variability.

In the subsequent discussion, the relevant statistical issues and the techniques designed to address them are illustrated using the specific example of *stacking* analysis. In such analyses, measurement data that correspond to multiple, independent but similar experiments are combined to enhance their scientific utility. In the context of γ-ray astronomy, stacking analyses are commonly applied in situations when a target population comprises a large number of individually undetectable source candidates that are theoretically expected to exhibit at least one identical signal characteristic. In the absence of genuine γ-ray emission, the superimposed observational data will be consistent with random Poisson fluctuations about the mean background level. Conversely, multiple faint signals produced by a genuine sample of faint γ-ray sources will reinforce upon combination, yielding a significant overall excess with respect to the measured background. Accordingly, stacking analyses often facilitate the derivation of average source properties for the target population as a whole, even when detection of individual γ-ray signals is rendered impossible by inadequate instrumental sensitivity. The specific consideration of stacking is instructive

because combined source analysis inherently segregates subsets of observational data which are likely to incorporate distinct systematic biases. Moreover, stacking analyses provide a more intuitive distinction between systematic effects that undergo temporal variation, such as telescope elevation or optical efficiency, and those that vary on a target-wise basis, such as the ambient night-sky brightness of the number of γ-ray sources within the telescope field-of-view.

Compelling and theoretically well motivated target populations for combined source analyses include X-ray binary systems (Dickinson, 2009), combined pulsar populations within globular clusters (Aharonian et al., 2009) and the self-annihilating dark matter components of dwarf spheroidal galaxies (e.g. H. E. S. S. Collaboration et al., 2011).

1.1. Relevant Aspects of the Observational Technique

The atmosphere of the Earth is almost completely opaque to multi-GeV and TeV radiation and accordingly, the operational mode of IACT arrays relies upon indirect detection of astrophysical γ-rays. Indeed, the majority of incident VHE photons undergo electron-positron pair production interactions within the Coulomb fields of stratospheric nuclei. Frequently, these interactions initiate electromagnetic cascades of charged particles, which radiate Cherenkov light as they descend superluminally through the dielectric medium of the atmosphere. IACT arrays are designed to image the resultant distribution of Cherenkov photons at ground-level, using the encoded information to reconstruct the properties of the progenitor γ-ray. The set of reconstructed γ-ray properties define a multi-dimensional parameter space, with suitable sub-spaces that define the event populations that are required for specific astrophysical analyses.

The situation is complicated somewhat by the fact that IACT arrays operate in a strongly background dominated regime and must successfully distinguish a comparatively weak γ-ray signal from an almost overwhelming background of air-showers initiated by hadronic cosmic rays. Fortunately, the properties of hadron- and γ-ray-initiated air-showers are sufficiently disparate that an effective segregation is possible based on the characteristics of the captured Cherenkov images. Traditionally, a simplified elliptical parameterisation of the Cherenkov light distribution was used in conjunction with lookup tables of simulated air-shower properties to separate the signal and background event populations (Cawley et al., 1985; Hillas, 1985; Weekes et al., 1989). Modern computational resources permit utilisation of the complete Cherenkov images (e.g Le Bohec et al., 1998; de Naurois and Rolland, 2009) and the application of advanced machine-learning algorithms (e.g Fiasson et al., 2010) to further enhance the background rejection efficiency.

Despite these sophisticated filtering techniques, the sheer number of air-showers that trigger an IACT array during a given observation inevitably results in a subset of hadron-originated Cherenkov images which are sufficiently γ-ray-like in appearance to survive the event selection procedure. Quantitatively, the probability that a γ-ray-like event having specific reconstructed parameters will pass background rejection is described by a multi-dimensional *acceptance* function. The reconstructed γ-ray signal (N_{ON}) within a suitably defined sampling

region of the event property parameter space is a superposition of the true γ-ray signal (N_S) on an irreducible background (N_B) of spuriously classified events.

$$N_\mathrm{ON} = N_\mathrm{S} + N_\mathrm{B} \tag{1}$$

The magnitude of the background component is typically estimated using the number of γ-ray-like events (N_OFF) that are reconstructed within one or more nominally OFF-source regions. Adopting this convention, (1) is normally re-expressed as

$$N_\mathrm{ON} = N_\mathrm{S} + \alpha N_\mathrm{OFF} \tag{2}$$

where α is a normalisation factor which compensates for any disparities in the instrumental response among the various sampling regions (see e.g. Berge et al., 2007, for a more detailed discussion).

Fundamentally, the precise value of α is dependent upon specific characteristics of the individual observations that contribute to the overall data set. Numerous variable factors such as the target zenith and azimuthal angles, the configuration and functionality of individual telescopes within a larger array, or the ambient atmospheric conditions may modify the nominal system performance.

The functional dependency of α on the various observational and instrumental parameters is complicated and is must typically be resolved for each observation using a multi-dimensional model estimate of the prevailing system acceptance (Berge et al., 2007). Candidate models may be synthesised by interpolating the distributions of representative γ-ray-like background data sets, assembled from multiple independent observations of empty fields-of-view. Accurate replication of the true system acceptance using this technique requires an extensive database of suitable observations with characteristics that sample the entire space of observational parameters with adequate resolution. For certain well-understood parameters it may be possible to derive semi-analytic parameterisations of the corresponding systematic distortions, which can improve the interpolation accuracy and reduce the required sampling density. Nonetheless, for complicated instruments with a large number of possible configurations, the observation time required to populate the reference database may be prohibitive. Alternatively, response models which inherently incorporate the appropriate systematic offsets can be constructed on an observation-wise basis using the subset of γ-ray-like background events that fall outside of the fiducial sampling regions. However, this approach is susceptible to contamination by imperfectly excluded γ-ray sources and is ultimately limited by the availability of event statistics within a single field-of-view.

For the current generation of IACTs, residual discrepancies between the generated model and the true instrumental response reflect a superposition of imperfectly modelled systematic offsets and are typically at the few percent level (Berge et al., 2007). However, if observational situations arise in which both approaches for modelling the system acceptance are compromised, the resultant disparity could be much larger. For example, if the telescope configuration or the prevailing operating conditions restrict the availability of appropriate archival

background data and the target field-of-view is crowded, then systematic effects could dominate statistical uncertainties when searching for faint γ-ray sources. Moreover, since the combined impact of multiple systematic effects is often reflected by *non-uniform* variations in the system acceptance, the relative extent and location of the sampling regions within the event property parameter space may also bias the estimated value of α. Most experiments tacitly acknowledge the uncertainty in α by attaching a conservative systematic error term to their quoted results.

1.2. Alternative Stacking Analyses

Conventional stacking analyses typically entail simple summation of the observed event counts associated with the individual targets in the population sample, with the aggregate datasets forming the basis for subsequent astrophysical inference. However, the background estimation procedures outlined in §1.1 render this traditional approach (hereafter referred to as *data stacking*) suboptimal in the context of VHE γ-ray observations. In particular, data stacking analyses are complicated somewhat by the requirement to synthesise a combined value of the normalisation parameter $\bar{\alpha}$ which corresponds to the stacked dataset. A viable approach uses the definition of the overall signal excess,

$$
\begin{aligned}
\Delta &= N_{\mathrm{ON}} - \bar{\alpha} N_{\mathrm{OFF}} \\
&= \sum_i N_{\mathrm{ON},i} - \bar{\alpha} \sum_i N_{\mathrm{OFF},i}, \quad i = 1, \ldots, m \\
&= \sum_i N_{\mathrm{ON},i} - \sum_i \alpha_i N_{\mathrm{OFF},i}, \quad i = 1, \ldots, m
\end{aligned}
\tag{3}
$$

to form an average over all m targets in the sample, weighted by the individual off-source counts.

$$
\bar{\alpha} = \frac{\displaystyle\sum_i \alpha_i N_{\mathrm{OFF},i}}{\displaystyle\sum_i N_{\mathrm{OFF},i}}, \quad i = 1, \ldots, m
\tag{4}
$$

Unfortunately, this technique fails to account for any target-specific uncertainties that arise during derivation of the individual α_i. Even if no systematic *biases* are introduced, the diverse observational parameters that correspond to each individual target dataset will likely affect accuracy to which the corresponding values of α_i can be determined (Berge et al., 2007). The resultant non-uniformity renders subsequent propagation of the individual error estimates to the value of $\bar{\alpha}$ somewhat problematic for the data stacking approach. Indeed, it is often unclear how the individual uncertainties should be combined in a statistically consistent manner, particularly when they are characterised by asymmetric intervals or more complicated probability distributions. Assignment of a uniform but conservative systematic error term to all of the α_i renders the problem tractable, but inevitably results in an overall loss of sensitivity.

Moreover, indications exist that the shortcomings of the data stacking technique can be problematic in experimental situations. Indeed, Dickinson (2009)

derived anomalously high combined significance results when applying a data stacking procedure to two independent target ensembles comprising 37 and 44 putative VHE γ-ray sources for which $N_{\mathrm{OFF},i} \sim 100 - 500$ and $\alpha_i \sim 0.07$. By studying the ON and OFF-source event count distributions, Dickinson (2009) concluded that a relative target-wise dispersion $\Delta\alpha_i/\alpha_i \lesssim 0.1$ was required to explain the derived significances in the absence of a true γ-ray signal.

In subsequent sections, the statistical behaviour of the data stacking approach is investigated, highlighting the limited applicability of this technique in scenarios for which α is uncertain. In response to these inherent shortcomings, a statistically robust method for the combination of IACT observations is outlined, which allows systematic uncertainties to be treated on a target-by-target basis. Using this technique, observations yielding weakly constrained values of α_i may be usefully and consistently combined with more reliable datasets, without diluting the scientific utility of the latter.

2. Derivation of the Joint Likelihood

The new method operates by defining target-specific likelihood functions which facilitate appropriate treatment of relevant systematic uncertainties. Stacking is implemented by forming a product of individual target likelihood functions and using this combined likelihood to estimate the shared characteristics of the putative γ-ray signals. It should be emphasised that the applicability of the method is contingent upon an implicit assumption of at least one identical signal characteristic for all targets comprising the sample. Realistically, signal characteristics for which the required assumption of universality is reasonable are typically target-specific functions of the experimental observables. As discussed in the subsequent paragraphs, disparities between putative subpopulations comprising the target sample are inherently problematic in the context of data stacking. In contrast, the suggested approach intrinsically facilitates straightforward customisation of the individual target likelihoods.

Construction of the single target likelihood function proceeds under the assumption that the sample of γ-ray-like events that are reconstructed within a nominal sampling region is drawn from a parent population with Poisson distributed arrival times. Accordingly, if the value of α were precisely constrained, then (2) would imply

$$N_{\mathrm{ON}} \sim \mathrm{Pois}(N_{\mathrm{ON}}, \tilde{N}_{\mathrm{S}} + \tilde{\alpha}\tilde{N}_{\mathrm{OFF}}) \text{ and } N_{\mathrm{OFF}} \sim \mathrm{Pois}(N_{\mathrm{OFF}}, \tilde{N}_{\mathrm{OFF}}) \qquad (5)$$

where $\tilde{N}_{\mathrm{ON}}$ and $\tilde{N}_{\mathrm{OFF}}$ denotes the true mean value of the Poisson distributed variables N_{ON} and N_{OFF} respectively. The new parameter $\tilde{\alpha}$ represents the unknown, true value of the derived normalisation parameter. In fact, the assumed validity of (5) is implicit in the derivation of $\bar{\alpha}$ in (4) with likely deviations from this idealised behaviour undermining the reliability of the traditional data stacking framework. Conversely, the combined likelihood approach treats target-specific systematic uncertainties by modelling α as a random variable with arbitrary distribution $G(\alpha|\tilde{\alpha})$. As indicated in §1.1, the specific functional

definitions of $G(\alpha|\tilde{\alpha})$ are dependent upon observational parameters which are likely to vary on a target-by-target basis.

The probability density function which models an abstract single-target dataset $D = \{N_{\mathrm{ON}}, N_{\mathrm{OFF}}, \alpha\}$ is simply the product of the distributions of its component data.

$$
\begin{aligned}
f_1(N_{\mathrm{ON}}, N_{\mathrm{OFF}}, \alpha|\tilde{N}_{\mathrm{S}}, \tilde{N}_{\mathrm{OFF}}, \tilde{\alpha}) = {} & \mathrm{Pois}(N_{\mathrm{ON}}, \tilde{N}_{\mathrm{S}} + \tilde{\alpha}\tilde{N}_{\mathrm{OFF}}) \\
& \cdot \mathrm{Pois}(N_{\mathrm{OFF}}, \tilde{N}_{\mathrm{OFF}}) \\
& \cdot G(\alpha|\tilde{\alpha})
\end{aligned}
\tag{6}
$$

Accordingly, substitution for D using a concrete observational dataset $d = \{n_{\mathrm{ON}}, n_{\mathrm{OFF}}, \alpha_{\mathrm{obs}}\}$ enables calculation of the target-specific likelihood describing the true γ-ray signal.

$$
L_1(\tilde{N}_S, \tilde{N}_{\mathrm{OFF}}, \tilde{\alpha}|n_{\mathrm{ON}}, n_{\mathrm{OFF}}, \alpha_{\mathrm{obs}}) = f_1(n_{\mathrm{ON}}, n_{\mathrm{OFF}}, \alpha_{\mathrm{obs}}|\tilde{N}_{\mathrm{S}}, \tilde{N}_{\mathrm{OFF}}, \tilde{\alpha})
\tag{7}
$$

Modelling systematically uncertain parameters as random variables with known probability density functions is a common statistical approach, with applications extending beyond the limited domain of stacking analyses (e.g. Heinrich and Lyons, 2007). Indeed, equations (6) and (7) are equally valid in single target IACT analyses that must accommodate non-negligible systematic uncertainties on the estimated value of α.

The ultimate goal of this stacking procedure is to obtain improved constraints on the global γ-ray signal characteristics. In this context, $\tilde{N}_{\mathrm{OFF}}$ and $\tilde{\alpha}$ are not of primary interest and are therefore categorised as nuisance parameters throughout the subsequent analysis .

Finally, the joint likelihood describing the global signal characteristics of the entire ensemble of m targets is simply the product of the individual target likelihoods.

$$
L_{\mathrm{J},m}(\tilde{N}_{\mathrm{S}}) = \prod_{i=1}^{m} L_{1,i}(\tilde{N}_{\mathrm{S}}, \tilde{N}_{\mathrm{OFF},i}, \tilde{\alpha}_i|n_{\mathrm{ON},i}, n_{\mathrm{OFF},i}, \alpha_{\mathrm{obs},i})
\tag{8}
$$

Accordingly, the set of values which parameterise L_{J} is the union of the parameters of the component single-target likelihood functions and therefore incorporates $2m$ uncertain nuisance parameters.

3. Inference of the Signal Characteristics

3.1. Detection Significance

The initial objective of a stacking analysis is typically the conclusive detection of a combined γ-ray excess, with detailed investigation of specific signal characteristics assuming subsidiary importance. In such situations, the probability of detection is typically evaluated in terms of the the significance, S with which experimental data exclude the null hypothesis of zero signal.

For an abstract likelihood function $L(\boldsymbol{\pi}, \boldsymbol{\theta}|D)$ relating the experimental data $D = \{D_1 \ldots D_n\}$ to multiple nuisance parameters, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_j\}$ and parameters of interest $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_l\}$, a popular formula for calculation of detection significance (Li and Ma, 1983) employs the likelihood ratio defined by (Neyman and Pearson, 1933)

$$\lambda_{\mathrm{LM}} = \frac{L(\boldsymbol{\pi}_0, \hat{\boldsymbol{\theta}}_0|D)}{L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}|D)} \tag{9}$$

where a caret denotes the maximum likelihood estimate of the accented symbol, $\boldsymbol{\pi}_0$ denotes the value of $\boldsymbol{\pi}$ which corresponds to the null hypothesis, and $\hat{\boldsymbol{\theta}}_0$ represents the conditional maximum likelihood estimate for $\boldsymbol{\theta}$ given that $\boldsymbol{\pi} = \boldsymbol{\pi}_0$. The significance is subsequently defined in terms of λ_{LM}

$$S = \sqrt{-2 \ln \lambda_{\mathrm{LM}}} \tag{10}$$

which reduces to

$$S = \sqrt{2} \left\{ N_{\mathrm{ON}} \ln \left[\left(\frac{1+\alpha}{\alpha} \right) \frac{N_{\mathrm{ON}}}{N_{\mathrm{ON}} + N_{\mathrm{OFF}}} \right]^{N_{\mathrm{ON}}} + N_{\mathrm{OFF}} \ln \left[(1+\alpha) \frac{N_{\mathrm{OFF}}}{N_{\mathrm{ON}} + N_{\mathrm{OFF}}} \right]^{N_{\mathrm{ON}}} \right\}^{\frac{1}{2}} \tag{11}$$

as reported by (Li and Ma, 1983, Equation 17) and is nominally χ_1 distributed for a typical, single-target dataset with negligible α uncertainty i.e. $D = \{N_{\mathrm{ON}}, N_{\mathrm{OFF}}, \alpha\}$. Proponents of data stacking typically emulate the single-target approach by using Equation 17 in conjunction with a pseudo-dataset $D' = \{\sum_i N_{\mathrm{ON},i}, \sum_i N_{\mathrm{OFF},i}, \bar{\alpha}\}$. Consequently, non-uniform observational dependencies the individual α_i become entangled, which inevitably complicates statistically robust treatment of their effects. Conversely, direct substitution of L_J in the calculation of λ_{LM} in the joint-likelihood approach ensures that the resultant expression for S is constructed using all available information regarding target-specific systematic uncertainties.

Recently, Klepser (2011) proposed an extension to the method suggested by Li and Ma (1983) which uses well-calibrated estimates of the instrumental point-spread function to achieve enhanced sensitivity relative to the more traditional analysis. Their technique ameliorates the effect of systematic variations in the telescope acceptance characteristics by segregating the observational data on the basis of prevailing instrumental and environmental parameters.

3.2. Confidence Intervals

Results that are derived from a joint likelihood stacking analysis comprise estimates of the various experimental parameters of interest, with any uncertainty quantified using associated confidence intervals. The frequentist definition of a properly constructed $100(1 - \epsilon)\%$ confidence interval states that $100(1 - \epsilon)\%$ of such intervals, generated from a large number of independent experimental measurements of a quantity X, will contain (or *cover*) the true value $\tilde{X}$. Confidence intervals which fulfil this criterion are described as having correct

coverage. Construction of confidence intervals typically uses experimental data to define a test statistic which is subsequently used to identify regions of the relevant parameter space that fulfil the required statistical criteria. The *profile likelihood* is an appropriate choice of test statistic in situations that require the treatment of multiple nuisance parameters. The profile likelihood is defined as the ratio

$$\lambda_{\mathrm{PL}}(\boldsymbol{\pi}|D) = \frac{L(\boldsymbol{\pi}, \hat{\hat{\boldsymbol{\theta}}}_\pi|D)}{L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}|D)} \tag{12}$$

where previously encountered symbols retain their earlier definitions and $\hat{\hat{\boldsymbol{\theta}}}_\pi$ represents the conditional maximum likelihood estimate for $\boldsymbol{\theta}$ assuming a specific value of $\boldsymbol{\pi}$. Conveniently, the distribution of $-2 \log \lambda_{\mathrm{PL}}$ converges to that of a χ^2 random variate with l degrees of freedom for large experimental data sets (e.g. Casella and Berger, 2002). Accordingly, straightforward derivation of confidence intervals is facilitated by comparison of the derived test statistic with appropriate percentiles of the relevant χ^2 distribution. Moreover, the profile likelihood has the desirable property of being independent of $\boldsymbol{\theta}$, facilitating the derivation of robust confidence intervals, even in the presence of uncertain nuisance parameters.

4. Monte Carlo Studies

Verification of the performance and reliability joint likelihood stacking procedure employed a computerised toy Monte Carlo approach. The RooFit[2] framework (Verkerke and Kirkby, 2006) was used to generate multiple simulated datasets consistent with the assumed distributions of N_{ON}, N_{OFF} and α for specific values of the true signal parameters $\tilde{N}_{\mathrm{S}} \in [0, 10]$, $\tilde{N}_{\mathrm{OFF}} = 100$ and $\tilde{\alpha} = 0.1$. To better understand the effect of the distribution of α on the outcome of each stacking procedure, data were generated using three alternative parameterisations for $G(\alpha|\tilde{\alpha}_i)$. A basic test case, simulating unbiassed systematic uncertainties, employed a Gaussian parameterisation $G(\alpha|\tilde{\alpha}) = \mathrm{Gaus}(\tilde{\alpha}, \sigma_\alpha)$ with $\sigma_\alpha = 0.02$ (hereafter referred to as *Model A*). Two less idealised examples, adopting bifurcated Gaussian functions

$$G(\alpha|\tilde{\alpha}) = \begin{cases} \mathrm{Gaus}(\tilde{\alpha}, \sigma_{\alpha,\mathrm{L}}) & \text{if } \alpha < \tilde{\alpha} \\ \mathrm{Gaus}(\tilde{\alpha}, \sigma_{\alpha,\mathrm{R}}) & \text{if } \alpha > \tilde{\alpha} \end{cases} \tag{13}$$

with $(\sigma_{\alpha,\mathrm{L}}, \sigma_{\alpha,\mathrm{R}}) = (0.01, 0.03)$ and $(\sigma_{\alpha,\mathrm{L}}, \sigma_{\alpha,\mathrm{R}}) = (0.02, 0.08)$ (hereafter referred to as *Model B* and *Model C*, respectively), allowed the effect of increasing asymmetry in the modelled α distribution to be investigated. Figure 1 shows representative distributions of α corresponding to each of the simulated scenarios. Models A and B mimic somewhat pessimistic observational situations in which the choice of OFF-source sampling regions is restricted, limiting event
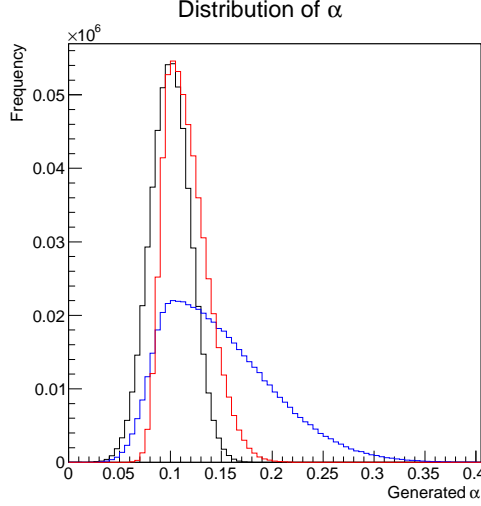
---

[2]http://roofit.sourceforge.net

Figure 1: Generated Monte Carlo distributions of the ON/OFF normalisation α corresponding to Model A (*black*), Model B (*red*), and Model C (*blue*).

statistics and biasing the derivation of α. Accordingly, the corresponding results are particularly relevant to galactic source populations, for which source confusion and diffuse background contamination within target fields-of-view can be particularly problematic. Model C represents an extreme case which is unlikely to occur in real experimental contexts, but is nonetheless included to demonstrate the efficacy of the joint likelihood technique.

Using RooFit, ensembles of the generated single-target datasets were then employed in conjunction with the relevant compound probability density functions to obtain multiple realisations of the joint likelihood $L_{J,m}$ corresponding to a discrete range of target multiplicities $m \in [1, 10]$. For each ensemble dataset D_m, the statistical significance (S_J) of the combined γ-ray signal was evaluated using the likelihood ratio prescription of (Li and Ma, 1983)

$$S_J = -2 \log \left\{ \frac{L_{J,m}(N_S = 0, \hat{\boldsymbol{\theta}}_0 | D_m)}{L_{J,m}(\hat{N}_S, \hat{\boldsymbol{\theta}} | D_m)} \right\}, \; \hat{\theta}_i = \{\hat{N}_{\text{OFF},i}, \hat{\alpha}_i\}, \; i = 1 \ldots m. \quad (14)$$

For comparison, a traditional data stacking analysis was also applied to the generated data. The components of the ensemble datasets used to construct each realisation of $L_{J,m}$ were combined to obtain corresponding stacked datasets, $D'_m = \{\sum_i N_{\text{ON},i}, \sum_i N_{\text{OFF},i}, \bar{\alpha}\}$. Subsequent statistical inference employed a likelihood function L_{DS}, which was formed by substitution of the stacked data into the single target probability density function f_1 with G defined as a delta function centred on the stacked $\bar{\alpha}$ estimate, $G(\alpha | \tilde{\alpha}, \bar{\alpha}) = \delta(\alpha - \bar{\alpha})$. Following the convention outlined in §3, Equation (11) was evaluated using the components

10

of D'_m do derive the statistical significance S_{DS} of the combined γ-ray signal corresponding to the data stacking approach.

For both stacking methodologies, construction of the associated profile likelihoods was accomplished using the intrinsic capabilities of RooFit, adopting $\tilde{N}_{\mathrm{S}}$ as the sole parameter of interest. Confidence intervals were then defined as the ranges of $\tilde{N}_{\mathrm{S}}$ outside which $-2\log\lambda_{\mathrm{PL}}$ exceeded an appropriate percentile of the nominal χ_1^2 distribution.

5. Comparison of the Alternative Stacking Techniques

For each distinct combination of the true signal parameters, modelled distribution of α, and target multiplicity, the Monte Carlo approach outlined in §4 was applied to derive 5000 independent estimates of N_S and S corresponding to each stacking technique. In combination with appropriate statistical metrics, the resultant data enable a comparative evaluation of the joint likelihood and traditional data stacking approaches.

5.1. Detection Significance

An experimentally relevant diagnostic for both stacking analyses is the probability with which data corresponding to a particular value of $\tilde{N}_{\mathrm{S}}$ will yield a significance S, in excess of a specified threshold S_{t}. Assuming data that are distributed in accordance with the null hypothesis, this probability is called the *significance level*. Although the specific choice of S_{t} is arbitrary, it is typically chosen in order to obtain a desired significance level and ensure a controlled rate of false-positive detections. Objective comparison of the alternative stacking techniques is facilitated by a threshold which corresponds to a nominal percentile of the appropriate null distribution of S. A related statistic is the *power*, which describes the probability that derived values of $S > S_{\mathrm{t}}$ will correctly identify a genuine signal, and provides useful insights regarding the sensitivity of the corresponding analytical method.

The relevant Monte Carlo distributions for S, corresponding to $\tilde{N}_{\mathrm{S}} = 0$, are plotted for each target multiplicity in Figure 2. Assuming that the true value of $\tilde{\alpha}$ is 0.1, systematic uncertainties which affect the estimation of α_i according to the distributions plotted in Figure 1 will generate misleading results in a data stacking analysis. In many experimental situations, a small number of erroneous results is tolerable, as long as the expected frequency with which they occur can be reliably predicted. This predictability criterion is realised if the derived values consistently conform to a known statistical distribution. Irrespective of $G(\alpha|\tilde{\alpha})$, and for all $m \in [0, 10]$, the significance estimates derived using the joint likelihood technique appear to retain the χ_1 distribution which arises in case of zero α uncertainty. In contrast, results that correspond to traditional data stacking exhibit substantial deviations from this nominal distribution. Moreover, adoption of asymmetric α distributions introduces a monotonic dependence on the target multiplicity, yielding systematically larger significance estimates as m increases.

11

Consideration of the procedure used during data stacking to synthesise a combined α estimate suggests an intuitive explanation of the observed behaviour. Models B and C generate ensemble data sets in which the subset of overestimated α values is typically predominant and has an absolute *mean* offset from $\tilde{\alpha}$ which exceeds that of the remaining, underestimated α_i. Accordingly, the frequency with which Equation (4) yields a substantial overestimate for $\bar{\alpha}$ should increase with the ensemble size, which is consistent with the observed behaviour.

Conversely, symmetrically distributed α uncertainties should not inherently bias the data stacking process. Instead, the target-specific weighting of each α_i specified in Equation (4) prevents perfect cancellation of the systematic offsets and inhibits convergence to the nominal null distribution for S as $m \rightarrow \infty$. As outlined in §4 all Monte Carlo data sets used for this study assume a common $\tilde{N}_{\mathrm{OFF}} = 100$, with the range of inter-target variability in $N_{\mathrm{OFF},i}$ restricted to the level of Poisson fluctuations. In experimental contexts, the true number of OFF events is likely to be different for each target and the weights accorded to the corresponding α_i would be more diverse. The adverse impact of symmetric systematic uncertainties would then depend critically on the joint, target-wise distribution of α-offset and $N_{\mathrm{OFF},i}$, with specific ensembles potentially realising particularly benign or pathological scenarios.

As an alternative to the absolute measure of significance defined by Equation (11), experimental practitioners often choose to multiply S by the sign of the corresponding γ-ray excess. Adopting this convention, systematic *over*estimates of $\bar{\alpha}$ during data stacking would tend to over-generate negative significances when $\tilde{N}_{\mathrm{S}} = 0$. In many situations the notion of a negative signal is not physically meaningful and such results would probably be disregarded. In general, the ability to eliminate spurious results on the basis of their physical validity may ameliorate the impact of particular systematic effects, and reduce the rate of false-positive detections. Nonetheless, unfeasibly negative significances should raise concerns regarding the trustworthiness of corresponding upper limits.

Significance thresholds corresponding to the 95th percentile of the various null distributions of S are plotted in the left-hand column of Figure 3. Results derived using the joint-likelihood approach exhibit minimal multiplicity dependence and appear consistent with the nominal $S_t = 1.96$ expected for a χ_1 random variate. Conversely, thresholds which correspond to the data stacking approach systematically exceed their joint likelihood counterparts, and become increasing discrepant for larger m in simulations that assume asymmetric distributions of α. Accordingly, recourse to a nominal distribution of S is not appropriate in the data stacking framework, and valid interpretation of the significance of a particular detection requires empirical calibration of distinct significance thresholds for each target multiplicity. The values of S_t plotted in Figure 3 permit derivation of appropriately calibrated estimates of the statistical power at each vertex of the investigated model parameter space. The resultant probabilities are presented in Figure 4, while the results plotted in the right-hand column of Figure 3 reveal the corresponding differences power between the alternative stacking techniques.

12

Results pertaining to Models A and B exhibit saturation at extreme values of $\tilde{N}_S$ and m, caused by the limited statistics of the Monte Carlo datasets. Results corresponding a symmetric Gaussian distribution for α reveal derived powers which are compatible to $< 2\%$ for all $m \in [1, 10]$ and $\tilde{N}_S \in [0, 10]$. Although traditional data stacking appears to offer marginally superior performance for this scenario, asymmetric parameterisations for the distribution of α reverse this outcome. Indeed, powers derived using the joint likelihood approach in conjunction with Models B and C exhibit substantial enhancement with respect to their data stacking counterparts. Moreover, powers corresponding to the data stacking technique exhibit a marked sensitivity to the degree of asymmetry in the adopted α distribution, which appears to be substantially ameliorated by the joint likelihood approach. Strong suppression of the resultant data stacking power is evident when Model C is applied to the Monte Carlo analysis, with derived discrepancies ~ 0.9 separating the alternative stacking techniques at the upper extremes of the modelled parameter ranges. More significantly for practical applications, assuming the intermediate scenario represented by Model B, reveals a maximum disparity of ~ 0.4 for the experimentally interesting region of $\tilde{N}_S \sim 5$ as $m \to 10$. In this context, the joint likelihood approach yields a two-fold enhancement in the likelihood of detecting a genuine γ-ray signal.

5.2. Confidence Intervals

For an analysis to perform reliably in an experimental context, it is important that any generated confidence intervals have correct coverage. In reality, despite sophisticated treatment of associated systematic and statistical uncertainties, residual disparities between the assumed and actual distributions of experimental data often prevent practical realisation of this criterion. Moreover, deviations from correct coverage often lead to an unexpected increase in the frequency of spuriously inferred scientific conclusions. Accordingly, the degree by which the derived confidence intervals deviate from their nominal coverage provides an informative comparator for the alternative stacking techniques.

Figure 5 displays the fraction of derived 95% confidence intervals for N_S which cover the true signal value, for each parameterisation of $G(\alpha|\tilde{\alpha})$. Complementary insight is provided by Figure 6, which illustrates the corresponding deviations from correct coverage as a function of $\tilde{N}_S$ and the target multiplicity, m. It is evident that confidence intervals generated by the traditional data stacking approach are generally incompatible with their stated coverage. While the derived coverage fractions exceed the nominal 95% for low signals and target multiplicities, the majority of the $(\tilde{N}_S, m)$ parameter space is characterised significant negative deviations. This behaviour is typically described as *under-coverage*, and implies excessively permissive confidence intervals with an increased propensity to generate false-positive results. The observed undercoverage is maximised for small non-zero signals and appears to be exacerbated by the addition of targets to the ensemble dataset. Moreover, increased asymmetry of the modelled α distribution has a markedly detrimental effect on the derived coverage characteristics. Indeed, the maximum divergence from correct coverage increases from 4% to 25% and then to 85% for Model A, Model B,

13

and Model C, respectively. In the worst case, this means that only 10% of the derived intervals include the true parameter value. The inherent inability of the data stacking technique to acknowledge systematic uncertainties leads to confidence intervals which are unrealistically narrow. Consequently, reported accuracy of experimental measurements will be excessively optimistic and in the worst cases could be used to infer spurious scientific conclusions.

In contrast, notable stability in the coverage fractions derived using the joint likelihood technique indicates robust handling of systematic uncertainties. Confidence intervals derived under the assumption of a Gaussian distribution in α exhibit coverage which is correct to within $\lesssim 1\%$ throughout the studied region of the model parameter space. Variation of the coverage fraction as a function of $\tilde{N}_{\mathrm{S}}$ remains minimal when asymmetric parametrisations of $G(\alpha|\tilde{\alpha})$ are assumed. Although under-coverage which worsens with increasing target multiplicity is evident in results that correspond to Models B and C, the discrepancy with respect to correct coverage does not exceed $\sim 5\%$.

6. Discussion

The comparative evaluation presented in §5 reveals several limitations to the applicability of traditional data stacking in situations when α is uncertain, many of which are addressed by the joint likelihood technique.

Reliable inference of signal characteristics is impaired by significant levels of under-coverage for $\tilde{N}_{\mathrm{S}} \gtrsim 2$ in the data stacking approach. Moreover, the inherent permissivity of derived confidence intervals is compounded by the inclusion of additional targets. In combination, these characteristics imply a substantially increased tendency to generate spurious results which is exacerbated as the stacked target ensemble is enlarged. In an experimental context, this behaviour would be highly undesirable and largely undermines the motivation for combined source analysis. In contrast, confidence intervals provided by the joint likelihood approach closely approximate correct coverage under the assumption of symmetric uncertainties on α, and exhibit multiplicity-dependent under-coverage at $\lesssim 5\%$ levels if an asymmetric distribution is adopted.

The deviations from correct coverage that are exhibited by the data stacking approach are consistent with the erroneous assumption of negligible uncertainty in the value of α. Resolution of this issue is complicated by the lack of a straightforward prescription for propagating systematic uncertainties to the calculated value of $\bar{\alpha}$. Conversely, the joint likelihood method provides a robust technique for the treatment of uncertain α estimates on a target-specific basis.

Appropriate interpretation of the combined signal significance is also more complicated within the data stacking framework. The Monte Carlo null distributions for S presented in Figure 2 indicate that application of Equation (11) requires proper calibration of S_t to prevent spurious rejection of the null hypothesis. The situation deteriorates when increasingly asymmetric parameterisations of the generated α distribution are adopted. Under these circumstances, naïve assumption of a nominal χ_1 distribution for S induces a systematic over-estimation of corresponding significance levels which worsens with increasing

target multiplicity. In an experimental context, the implied increase in the false-positive detection rate for larger target samples is contrary to common expectation and therefore represents a serious shortcoming of the data stacking technique. In contrast, significance estimates derived for $\tilde{N}_S$ using the joint likelihood approach appear insensitive to the value of m and display excellent correspondence with the expected χ_1 distribution.

Notwithstanding appropriate calibration of a suitable detection threshold S_t, the powers which characterise the data stacking technique indicate that asymmetry in the adopted α distribution substantially diminishes the overall sensitivity. Conversely, the powers corresponding to the joint likelihood analysis appear relatively unaffected by the choice of $G(\alpha|\tilde{\alpha})$. The alternative approaches exhibit similar properties, and indeed data stacking appears to marginally outperform the joint likelihood approach, if symmetric systematic uncertainties are assumed.

It should be acknowledged that the degree of adverse behaviour inherent in the data stacking technique is contingent upon the adopted values of $\tilde{\alpha}$ and $\tilde{N}_{\mathrm{OFF}}$, as well as the assumed magnitude of the simulated systematic effects. Accurate modelling of the system acceptance may ameliorate the observed pathologies, with the cumulative effect of target-specific α uncertainties only becoming significant for target ensembles with $m \gg 10$. Figure 7 illustrates the combined effect on data stacking of systematic offsets at the few percent level using calibrated values of the significance threshold that correspond to the 95th percentile of simulated null distributions for S. The Monte Carlo data sets assume relative target-wise uncertainties $0.01 \leq \sigma_{\alpha,i}/\tilde{\alpha} \leq 0.1$, for both symmetric and asymmetric parameterisations for $G(\alpha|\tilde{\alpha})$. The curves correspond to simulated ensembles comprising $m = 50$ target sources for which $\tilde{\alpha} = 0.1$ and $\tilde{N}_{\mathrm{OFF}} = 100$. Despite an inevitable reduction in the influence of the modelled systematic offsets, it is clear that even small target-specific biasses can reinforce to distort the combined significance for large observational datasets. Moreover, even the pessimistic parameter values that define Models A and B may be practically relevant, especially when the extent of OFF-source regions is restricted by nearby sources, or stringent event selection limits the overall count statistics.

The joint likelihood method demonstrates good performance in a number of simple but experimentally relevant scenarios where the reliability of traditional data stacking is seriously impaired. Nonetheless, it should be stated that the *full* potential of the joint likelihood method can only be realised if the distributions of all the systematic uncertainties affect each observational dataset are parameterisable with sufficient accuracy.

Admittedly, definitive derivation of the distribution of α is unlikely to be straightforward in many situations. Semi-analytic estimation of the required distributions necessitates detailed knowledge of the instrumental response as well as careful measurement of the environmental conditions at the time of observation. Accordingly, imperfect or incomplete understanding of the dominant systematic biases might render this approach untenable. Plausible alternative strategies include jackknife or bootstrap resampling of OFF-source data from randomly selected background sampling regions to derive multiple estimates

for α and construct an empirical distribution. In this case, the accuracy of the resultant parameterisation is limited by the finite number of permutations available for the resampling process. Emulating the approach of Klepser (2011) and segregating data on the basis of quantifiable *operating conditions* would enable the use of large representative datasets to derive template distributions that pertain to finite subsets of the observational parameter space.

In situations where the suggested techniques are unable to accurately reconstruct the shape of the α distribution, a generic parameterisation could be tailored using target-specific estimates of the dispersion in derived values of the ON/OFF normalisation. By acknowledging the existence of the associated systematic uncertainties, this first-order approximation would ameliorate under-coverage as well as deviations from the nominal χ_1 distribution for S, thereby reducing the rate of unexpected false-positive detections. Moreover, by facilitating limited adaptation of the modelled systematic uncertainties, this approach retains a key advantage of the full joint likelihood technique that cannot be straightforwardly emulated by traditional data stacking. Indeed, it is likely that this simplified approach might recover a significant proportion of the power which is achieved when the true distribution of α is available. Nonetheless, as demonstrated by the Monte Carlo results for a symmetric distribution of α, there may be situations in which the advantages offered by the joint likelihood approach are marginal and the performance of the data stacking approach is deemed sufficient.

While this study has focussed upon parameterisation of α uncertainties, the joint likelihood technique represents a highly flexible and extensible analytical approach. Indeed, related analyses are able to parameterise arbitrary properties of the instrument response, the measured γ-ray signal, and even suspected source properties such as the spectral index and temporal variability (e.g. Ackermann et al., 2011).

7. Conclusions

Toy Monte Carlo simulations have been used to investigate the effect of non-negligible systematic uncertainties affecting the relative normalisation of the ON and OFF-source event sampling regions in atmospheric Cherenkov telescope observations. As a specific example, the properties of two alternative strategies for combined source analysis have been examined. Situations have been identified in which the traditional *data stacking* approach exhibits unexpected and undesirable statistical behaviour. Even when correctly calibrated null distributions for the combined significance are employed, the sensitivity of the data stacking approach appears markedly impaired if asymmetrically distributed α uncertainties are assumed. The data stacking technique is also yields unreliable estimates for the γ-ray signal parameters, with derived confidence intervals deviate from their stated coverage by a margin which widens with increasing target multiplicity.

An alternative to data stacking has been outlined which combines target-specific likelihood functions that explicitly account for non-uniform systematic

uncertainties. As an analytical approach, the *joint likelihood* technique exhibits many characteristics which surpass those of traditional data stacking. It offers a statistically robust prescription for the treatment of target-specific systematic uncertainties, and effectively addresses the shortcomings which are inherent in the data stacking framework. Indeed, the joint likelihood effectively eliminates pathological behaviour inherent in data stacking, with combined source significances conforming to a single, nominal null distribution. Moreover, it achieves equivalent or superior sensitivity to the data stacking technique and yields parameter confidence intervals that exhibit minimal deviation from their stated coverage.

Many of the insights provided by this study of combined source analyses are equally applicable to single-target analyses involving similarly abundant event statistics. Indeed, the synthesis of any extensive dataset often combines observations which incorporate diverse systematic effects. This study has shown that the joint likelihood method provides a viable approach for the robust combination of data with inhomogeneous systematic uncertainties, irrespective of the celestial target coordinates. Although it is a generally applicable technique, the joint likelihood analysis is likely to prove most useful in experimental situations involving limited event statistics and a restricted choice of OFF-source sampling regions.

References

Ackermann, M., Ajello, M., Albert, A., Atwood, W. B., Baldini, L., Ballet, J., Barbiellini, G., Bastieri, D., Bechtol, K., Bellazzini, R., Berenji, B., Blandford, R. D., Bloom, E. D., Bonamente, E., Borgland, A. W., Bregeon, J., Brigida, M., Bruel, P., Buehler, R., Burnett, T. H., Buson, S., Caliandro, G. A., Cameron, R. A., Cañadas, B., Caraveo, P. A., Casandjian, J. M., Cecchi, C., Charles, E., Chekhtman, A., Chiang, J., Ciprini, S., Claus, R., Cohen-Tanugi, J., Conrad, J., Cutini, S., de Angelis, A., de Palma, F., Dermer, C. D., Digel, S. W., Do Couto E Silva, E., Drell, P. S., Drlica-Wagner, A., Falletti, L., Favuzzi, C., Fegan, S. J., Ferrara, E. C., Fukazawa, Y., Funk, S., Fusco, P., Gargano, F., Gasparrini, D., Gehrels, N., Germani, S., Giglietto, N., Giordano, F., Giroletti, M., Glanzman, T., Godfrey, G., Grenier, I. A., Guiriec, S., Gustafsson, M., Hadasch, D., Hayashida, M., Hays, E., Hughes, R. E., Jeltema, T. E., Jóhannesson, G., Johnson, R. P., Johnson, A. S., Kamae, T., Katagiri, H., Kataoka, J., Knödlseder, J., Kuss, M., Lande, J., Latronico, L., Lionetto, A. M., Llena Garde, M., Longo, F., Loparco, F., Lott, B., Lovellette, M. N., Lubrano, P., Madejski, G. M., Mazziotta, M. N., McEnery, J. E., Mehault, J., Michelson, P. F., Mitthumsiri, W., Mizuno, T., Monte, C., Monzani, M. E., Morselli, A., Moskalenko, I. V., Murgia, S., Naumann-Godo, M., Norris, J. P., Nuss, E., Ohsugi, T., Okumura, A., Omodei, N., Orlando, E., Ormes, J. F., Ozaki, M., Paneque, D., Parent, D., Pesce-Rollins, M., Pierbattista, M., Piron, F., Pivato, G., Porter, T. A., Profumo, S., Rainò, S., Razzano, M., Reimer, A., Reimer, O., Ritz, S., Roth, M., Sadrozinski, H. F.-W., Sbarra, C., Scargle, J. D., Schalk, T. L., Sgrò, C.,

Siskind, E. J., Spandre, G., Spinelli, P., Strigari, L., Suson, D. J., Tajima, H., Takahashi, H., Tanaka, T., Thayer, J. G., Thayer, J. B., Thompson, D. J., Tibaldo, L., Tinivella, M., Torres, D. F., Troja, E., Uchiyama, Y., Vandenbroucke, J., Vasileiou, V., Vianello, G., Vitale, V., Waite, A. P., Wang, P., Winer, B. L., Wood, K. S., Wood, M., Yang, Z., Zimmer, S., Kaplinghat, M., Martinez, G. D., Dec. 2011. Constraining Dark Matter Models from a Combined Analysis of Milky Way Satellites with the Fermi Large Area Telescope. Physical Review Letters 107 (24), 241302.

Aharonian, F., Akhperjanian, A. G., Anton, G., Barres de Almeida, U., Bazer-Bachi, A. R., Becherini, Y., Behera, B., Bernlöhr, K., Boisson, C., Bochow, A., Borrel, V., Braun, I., Brion, E., Brucker, J., Brun, P., Bühler, R., Bulik, T., Büsching, I., Boutelier, T., Chadwick, P. M., Charbonnier, A., Chaves, R. C. G., Cheesebrough, A., Chounet, L.-M., Clapson, A. C., Coignet, G., Dalton, M., Daniel, M. K., Davids, I. D., Degrange, B., Deil, C., Dickinson, H. J., Djannati-Ataï, A., Domainko, W., O'C. Drury, L., Dubois, F., Dubus, G., Dyks, J., Dyrda, M., Egberts, K., Emmanoulopoulos, D., Espigat, P., Farnier, C., Feinstein, F., Fiasson, A., Förster, A., Fontaine, G., Füßling, M., Gabici, S., Gallant, Y. A., Gérard, L., Giebels, B., Glicenstein, J. F., Glück, B., Goret, P., Hauser, D., Hauser, M., Heinz, S., Heinzelmann, G., Henri, G., Hermann, G., Hinton, J. A., Hoffmann, A., Hofmann, W., Holleran, M., Hoppe, S., Horns, D., Jacholkowska, A., de Jager, O. C., Jung, I., Katarzyński, K., Katz, U., Kaufmann, S., Kendziorra, E., Kerschhaggl, M., Khangulyan, D., Khélifi, B., Keogh, D., Komin, N., Kosack, K., Lamanna, G., Lenain, J.-P., Lohse, T., Marandon, V., Martin, J. M., Martineau-Huynh, O., Marcowith, A., Maurin, D., McComb, T. J. L., Medina, M. C., Moderski, R., Moulin, E., Naumann-Godo, M., de Naurois, M., Nedbal, D., Nekrassov, D., Niemiec, J., Nolan, S. J., Ohm, S., Olive, J.-F., de Oña Wilhelmi, E., Orford, K. J., Ostrowski, M., Panter, M., Paz Arribas, M., Pedaletti, G., Pelletier, G., Petrucci, P.-O., Pita, S., Pühlhofer, G., Punch, M., Quirrenbach, A., Raubenheimer, B. C., Raue, M., Rayner, S. M., Reimer, O., Renaud, M., Rieger, F., Ripken, J., Rob, L., Rosier-Lees, S., Rowell, G., Rudak, B., Rulten, C. B., Ruppel, J., Sahakian, V., Santangelo, A., Schlickeiser, R., Schöck, F. M., Schröder, R., Schwanke, U., Schwarzburg, S., Schwemmer, S., Shalchi, A., Skilton, J. L., Sol, H., Spangler, D., Stawarz, L., Steenkamp, R., Stegmann, C., Superina, G., Szostek, A., Tam, P. H., Tavernet, J.-P., Terrier, R., Tibolla, O., van Eldik, C., Vasileiadis, G., Venter, C., Venter, L., Vialle, J. P., Vincent, P., Vivier, M., Völk, H. J., Volpe, F., Wagner, S. J., Ward, M., Zdziarski, A. A., Zech, A., May 2009. HESS upper limit on the very high energy γ-ray emission from the globular cluster 47 Tucanae. A&A499, 273–277.

Berge, D., Funk, S., Hinton, J., May 2007. Background modelling in very-high-energy γ-ray astronomy. A&A466, 1219–1229.

Casella, G., Berger, R., 2002. Statistical inference. Duxbury advanced series in

statistics and decision sciences. Thomson Learning.
URL `http://books.google.se/books?id=0x_vAAAAMAAJ`

Cawley, M. F., Fegan, D. J., Gibbs, K., Gorham, P. W., Hillas, A. M., Lamb, R. C., Liebing, D. F., MacKeown, P. K., Porter, N. A., Stenger, V. J., Aug. 1985. Application of imaging to the atmospheric Cherenkov technique. In: Jones, F. C. (Ed.), International Cosmic Ray Conference. pp. 453–456.

de Naurois, M., Rolland, L., Dec. 2009. A high performance likelihood reconstruction of γ-rays for imaging atmospheric Cherenkov telescopes. Astroparticle Physics 32, 231–252.

Dickinson, H., 2009. Very high energy gamma-rays from binary systems. Ph.D. thesis, University of Durham.
URL `http://etheses.dur.ac.uk/290/1/Thesis.pdf`

Fiasson, A., Dubois, F., Lamanna, G., Masbou, J., Rosier-Lees, S., Aug. 2010. Optimization of multivariate analysis for IACT stereoscopic systems. Astroparticle Physics 34, 25–32.

H. E. S. S. Collaboration, Abramowski, A., Acero, F., Aharonian, F., Akhperjanian, A. G., Anton, G., Barnacka, A., Barres de Almeida, U., Bazer-Bachi, A. R., Becherini, Y., Becker, J., Behera, B., Bernlöhr, K., Bochow, A., Boisson, C., Bolmont, J., Bordas, P., Borrel, V., Brucker, J., Brun, F., Brun, P., Bulik, T., Büsching, I., Carrigan, S., Casanova, S., Cerruti, M., Chadwick, P. M., Charbonnier, A., Chaves, R. C. G., Cheesebrough, A., Chounet, L.-M., Clapson, A. C., Coignet, G., Conrad, J., Dalton, M., Daniel, M. K., Davids, I. D., Degrange, B., Deil, C., Dickinson, H. J., Djannati-Ataï, A., Domainko, W., Drury, L. O. C., Dubois, F., Dubus, G., Dyks, J., Dyrda, M., Egberts, K., Eger, P., Espigat, P., Fallon, L., Farnier, C., Fegan, S., Feinstein, F., Fernandes, M. V., Fiasson, A., Fontaine, G., Förster, A., Füßling, M., Gallant, Y. A., Gast, H., Gérard, L., Gerbig, D., Giebels, B., Glicenstein, J. F., Glück, B., Goret, P., Göring, D., Hague, J. D., Hampf, D., Hauser, M., Heinz, S., Heinzelmann, G., Henri, G., Hermann, G., Hinton, J. A., Hoffmann, A., Hofmann, W., Hofverberg, P., Horns, D., Jacholkowska, A., de Jager, O. C., Jahn, C., Jamrozy, M., Jung, I., Kastendieck, M. A., Katarzyński, K., Katz, U., Kaufmann, S., Keogh, D., Kerschhaggl, M., Khangulyan, D., Khélifi, B., Klochkov, D., Kluźniak, W., Kneiske, T., Komin, N., Kosack, K., Kossakowski, R., Laffon, H., Lamanna, G., Lennarz, D., Lohse, T., Lopatin, A., Lu, C.-C., Marandon, V., Marcowith, A., Masbou, J., Maurin, D., Maxted, N., McComb, T. J. L., Medina, M. C., Méhault, J., Moderski, R., Moulin, E., Naumann, C. L., Naumann-Godo, M., de Naurois, M., Nedbal, D., Nekrassov, D., Nguyen, N., Nicholas, B., Niemiec, J., Nolan, S. J., Ohm, S., Olive, J.-F., de Oña Wilhelmi, E., Opitz, B., Ostrowski, M., Panter, M., Paz Arribas, M., Pedaletti, G., Pelletier, G., Petrucci, P.-O., Pita, S., Pühlhofer, G., Punch, M., Quirrenbach, A., Raue, M., Rayner, S. M., Reimer, A., Reimer, O., Renaud, M., de Los Reyes, R., Rieger, F., Ripken, J., Rob, L., Rosier-Lees, S., Rowell, G., Rudak, B., Rulten, C. B., Ruppel, J., Ryde, F., Sahakian, V.,

Santangelo, A., Schlickeiser, R., Schöck, F. M., Schönwald, A., Schwanke, U., Schwarzburg, S., Schwemmer, S., Shalchi, A., Sikora, M., Skilton, J. L., Sol, H., Spengler, G., Stawarz, L., Steenkamp, R., Stegmann, C., Stinzing, F., Sushch, I., Szostek, A., Tavernet, J.-P., Terrier, R., Tibolla, O., Tluczykont, M., Valerius, K., van Eldik, C., Vasileiadis, G., Venter, C., Vialle, J. P., Viana, A., Vincent, P., Vivier, M., Völk, H. J., Volpe, F., Vorobiov, S., Vorster, M., Wagner, S. J., Ward, M., Wierzcholska, A., Zajczyk, A., Zdziarski, A. A., Zech, A., Zechlin, H.-S., H.E.S.S. Collaboration, Mar. 2011. H.E.S.S. constraints on dark matter annihilations towards the sculptor and carina dwarf galaxies. Astroparticle Physics 34, 608–616.

Heinrich, J., Lyons, L., 2007. Systematic errors. Ann.Rev.Nucl.Part.Sci. 57, 145–169.

Hillas, A. M., Aug. 1985. Cerenkov light images of EAS produced by primary gamma. NASA. Goddard Space Flight Center 19th Intern. Cosmic Ray Conf., Vol. 3 p 445-448 (SEE N85-34862 23-93) 3, 445–448.

Klepser, S., Dec. 2011. A generalized likelihood ratio test statistic for Cherenkov telescope data. ArXiv e-prints (Accepted for publication in Astroparticle Physics).

Le Bohec, S., Degrange, B., Punch, M., Barrau, A., Bazer-Bachi, R., Cabot, H., Chounet, L. M., Debiais, G., Dezalay, J. P., Djannati-Atai, A., Dumora, D., Espigat, P., Fabre, B., Fleury, P., Fontaine, G., George, R., Ghesquiere, C., Goret, P., Gouiffes, C., Grenier, I. A., Iacoucci, L., Malet, I., Meynadier, C., Munz, F., Palfrey, T. A., Pare, E., Pons, Y., Quebert, J., Ragan, K., Renault, C., Rivoal, M., Rob, L., Schovanek, P., Smith, D., Tavernet, J. P., Vrana, J., Oct. 1998. A new analysis method for very high definition imaging atmospheric Cherenkov telescopes as applied to the CAT telescope. Nuclear Instruments and Methods in Physics Research A 416, 425–437.

Li, T.-P., Ma, Y.-Q., Sep. 1983. Analysis methods for results in gamma-ray astronomy. ApJ272, 317–324.

Neyman, J., Pearson, E. S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231 (694-706), 289–337.
URL http://rsta.royalsocietypublishing.org/content/231/694-706/289.short

Verkerke, W., Kirkby, D., 2006. The RooFit Toolkit for Data Modeling. In: L. Lyons & M. Karagöz Ünel (Ed.), Statistical Problems in Particle Physics, Astrophysics and Cosmology. p. 186.

Weekes, T. C., Cawley, M. F., Fegan, D. J., Gibbs, K. G., Hillas, A. M., Kowk, P. W., Lamb, R. C., Lewis, D. A., Macomb, D., Porter, N. A., Reynolds, P. T., Vacanti, G., Jul. 1989. Observation of TeV gamma rays from the Crab nebula using the atmospheric Cerenkov imaging technique. ApJ342, 379–395.
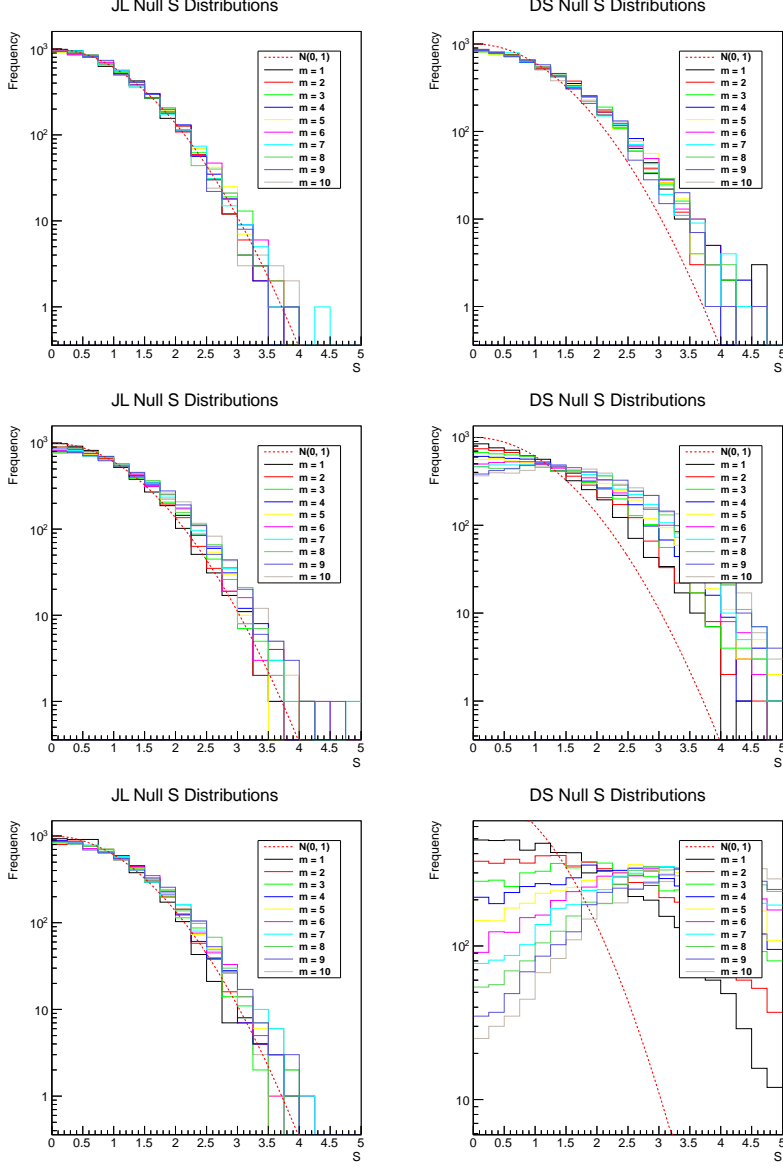
Figure 2: Monte Carlo distributions for S corresponding to the Joint Likelihood and Data Stacking approaches. Each histogram is representative of 5000 independent Monte Carlo datasets, which were generated assuming $\tilde{N}_S = 0$. Results are shown which correspond Model A (*top row*), Model B (*middle row*) and Model C (*bottom row*) for target multiplicities $m \in [1, 10]$. All Monte Carlo datasets were generated assuming $\tilde{N}_{\mathrm{OFF}} = 100$ and $\tilde{\alpha} = 0.1$.
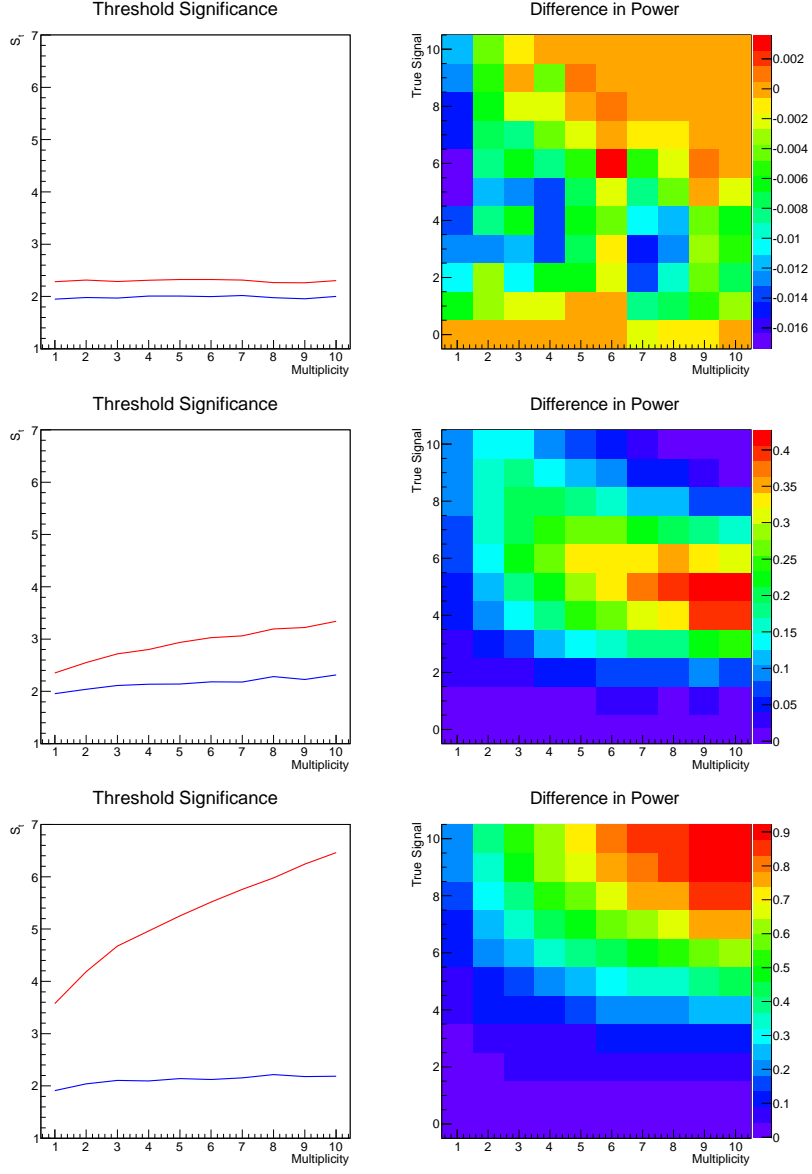
21

Figure 3: *Left-hand column:* Calibrated significance threshold values corresponding to the 95th percentile of the Monte Carlo null distributions of S for the Joint Likelihood (*blue*) and Data Stacking (*red*) approaches. *Right-hand column:* Differences between the derived powers corresponding to the Joint Likelihood and Data Stacking approaches. The plotted results correspond to Model A (*top row*), Model B (*middle row*) and Model C (*bottom row*). Each bin is derived using analyses of 5000 independent Monte Carlo datasets which assume to true signals $\bar{N}_{\rm S} \in [0, 10]$ and target multiplicities $m \in [1, 10]$. All Monte Carlo datasets were generated assuming $\tilde{N}_{\rm OFF} = 100$ and $\tilde{\alpha} = 0.1$.
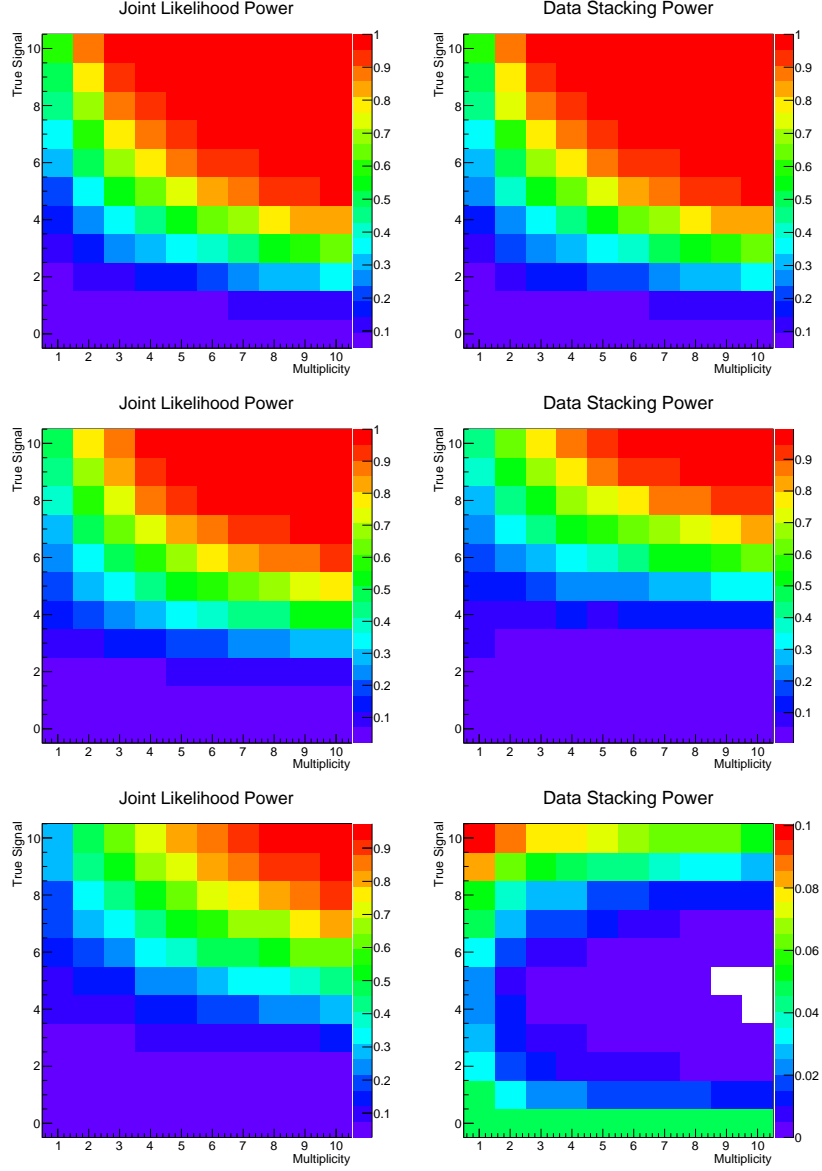
22

Figure 4: Statistical powers derived using S_t corresponding to the 95th percentile of the null distributions of S for the Joint Likelihood (*left-hand column*) and Data Stacking (*right-hand column*) approaches. Each bin is derived using analyses of 5000 independent Monte Carlo datasets. The plotted results correspond to Model A (*top row*), Model B (*middle row*) and Model C (*bottom row*) assuming true signals $\tilde{N}_S \in [0, 10]$ and target multiplicities $m \in [1, 10]$. The underlying datasets are as for Figure 3
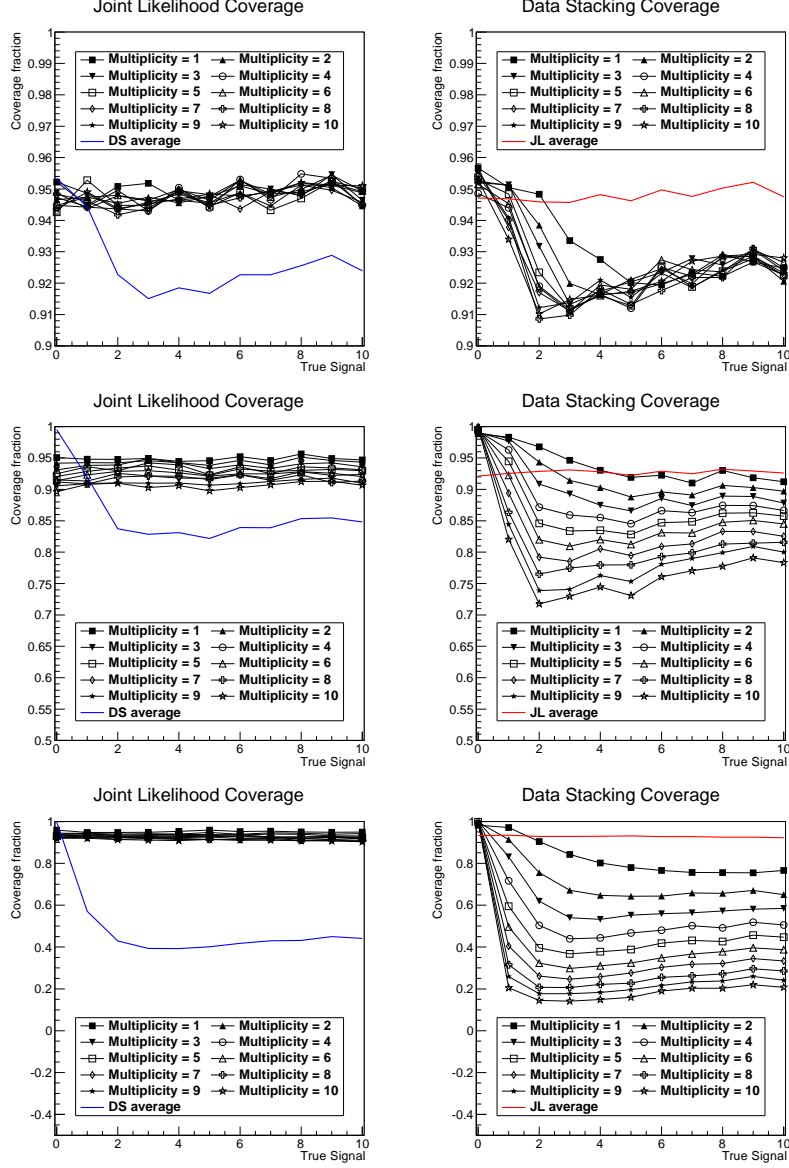
Figure 5: Joint Likelihood (JL) and Data Stacking (DS) coverage characteristics corresponding Model A (*top row*), Model B (*middle row*) and Model C (*bottom row*). Each datapoint is derived using analyses of 5000 independent Monte Carlo datasets and indicates the fraction of generated 95% confidence intervals on N_S which include the true signal value, $\tilde{N}_S$. The underlying datasets are as for Figure 3.
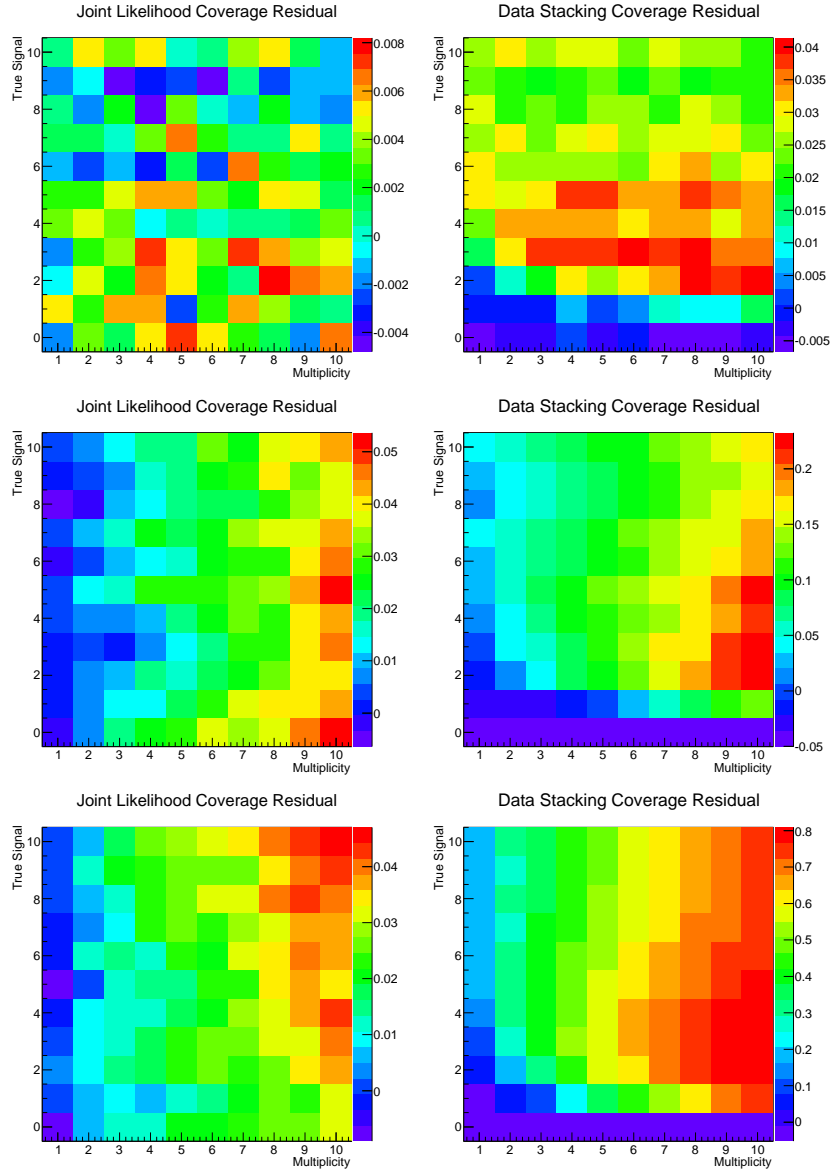
24

Figure 6: Residuals with respect to nominal coverage for 95% confidence intervals on N_{S} corresponding to the Joint Likelihood (JL) and Data Stacking (DS) approaches. Results are shown which correspond Model A (*top row*), Model B (*middle row*) and Model C (*bottom row*). The underlying datasets are as for Figure 3.
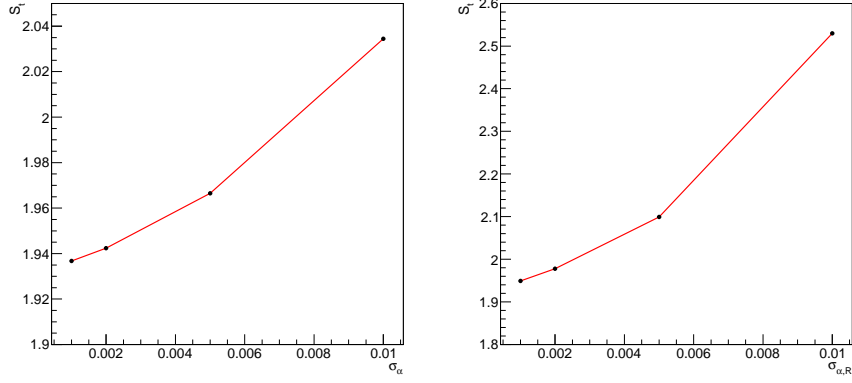
Figure 7: Calibrated significance threshold values corresponding to the 95th percentile of the Monte Carlo null distributions of S for the and Data Stacking approach. The illustrated results correspond to Gaussian (*left*) and bifurcated Gaussian (*right*) parameterisations for $G(\alpha|\tilde{\alpha})$, using target-specific uncertainties $1\% \leq \sigma_{\alpha}/\tilde{\alpha} \leq 10\%$ for the symmetric models, and $1\% \leq \sigma_{\alpha,\mathrm{R}}/\tilde{\alpha} \leq 10\%$ with $\sigma_{\alpha,\mathrm{R}} = 2\sigma_{\alpha,\mathrm{L}}$ for the asymmetric models. Each datapoint is derived using analyses of 5000 independent Monte Carlo datasets, which assume ensembles of $m = 50$ targets having true signals $\tilde{N}_{\mathrm{S}} = 0$. All Monte Carlo datasets were generated assuming $\tilde{N}_{\mathrm{OFF}} = 100$ and $\tilde{\alpha} = 0.1$.